Navigating Renewable Energy Markets

A Utility Company's Guide to Effective Forecasting

MGT 6203 Group Project Final Report

[Team 37 GitHub Page]

Nov 30, 2023

Amanda Wijntjes, Carina Grady, Josh Feldman, Micah Owens, Ty Underwood

Table of Contents

OBJECTIVE

We aim to predict electric demand and pricing by analyzing weather patterns and energy generation options so that utility companies can make informed decisions on renewable energy investments. With the increasing global emphasis on sustainable energy, utility companies are under pressure to invest more in renewable energy sources and optimize their current operations. Understanding energy consumption patterns, generation capabilities, and the influence of external factors like weather can help in making informed decisions. Spain has been at the forefront of renewable energy, with a significant portion of its energy consumption coming from renewable sources. With the increasing variability of intermittent renewable energy sources, such as wind and solar, there's a need to understand and predict energy consumption patterns and pricing to optimize operations and investments. A Transmission System Operator (TSO) oversees the high-voltage transmission grid, ensuring electricity transfer from generation sources to consumers. The dataset, sourced from ENTSOE and Spanish TSO Red Electric España, offers hourly electrical consumption and forecast data, complemented by weather insights from the Open Weather API for Spain's major cities.

Improving our forecasting capabilities for electrical consumption and pricing presents an exciting proposition from a business perspective. Accurate forecasts allow for optimized resource allocation, reducing the need for costly last-minute adjustments and minimizing excess capacity. This operational efficiency directly translates to substantial cost savings. From a customer perspective, consistent energy supply, fewer outages, and transparent pricing enhances consumer satisfaction and confidence in the company. Predicting and optimizing peak load facilitates the seamless integration of renewable energy sources into our grid. This foresight is invaluable, allowing us to prepare for shifts in energy goals and ensuring our continued growth and relevance in a rapidly evolving market.

Research Questions

- How do energy consumption patterns and pricing correlate with weather patterns, and how can this information be used to predict future energy pricing and demand?
- How well does our predictive model perform in forecasting electrical pricing, and can we demonstrate improved accuracy compared to TSO forecasts, particularly when considering time-of-day variations?
- Which specific weather measurements (e.g., temperature, humidity, wind speed) have the most significant impact on electrical consumption and pricing within the dataset?
- How can we optimize the utility's renewable energy operations based on predicted energy demands?

HYPOTHESIS

We anticipate clear patterns will emerge in the data to indicate one or more of the analyzed predictors have a significant influence on electrical demand, prices, and generation capacity. We expect that weather measurements will have a significant impact on electrical demand and pricing, which can inform the energy company's decisions and budget for green energy investment. It is expected that our decision tree approach and error verification using out-of-bag sampling will provide a more accurate and reliable predictive model for electrical consumption compared to the Transmission System Operator forecasts.

METHODOLOGY

Load and cost are two critical factors we wish to examine in this study. We feel that a random forest is a good choice for load prediction due to non-linear relationships between features like time of day and load. The data are auto-correlated so randomly splitting the data would not be ideal for this model. Since the dataset spans Jan 1, 2015 through Feb 26, 2016, we have one complete year of data for our study. The 2016 dates are left for testing. Out-of-bag error will be used for verification of our model throughout each permutation of the training phase. Feature selection will be decided after building a model with all of our independent variables. Only features responsible for large variance reduction will be selected for the remainder of model training. The number of estimators and number of features to consider at a split can then be "optimized" as the value at which out-of-bag error levels off.

While understanding load patterns is crucial for operations efficiency and reliability, it is important for us to understand electricity cost for business growth and renewable investment. We will also examine the relationship between various energy generation methods and cost using a simple linear regression model. We will test a full model as well as several variable selection methods. Assuming we find these sources significant, R^2 can be used to interpret how much of the price can be explained by these energy generation sources. We have also already observed interactions with features like time of day, temperature, and precipitation, as peak demand times and inclement weather may naturally lead to increased demand.

DATA OVERVIEW

Our dataset includes four years of electrical consumption, generation, pricing, and weather data for Spain. The weather dataset contains 17 columns of data: datetime index localized to CET, city, temperature (Kelvin), min temperature (Kelvin), max temperature (Kelvin), pressure (hPa), humidity (%), wind speed (m/s), wind direction (degrees), rain - last hr (mm), rain - last 3 hrs (mm), snow - last 3 hrs (mm), cloud cover (%), weather description code, short current weather description, long current weather description, weather icon code. Accessing the data can be done through the following source: Hourly energy demand generation and weather (kaggle.com)

Table 1. Sample data rows from the weather dataset.

1	dt_iso	city_name	temp	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	rain_1h	rain_3h	snow_3h	clouds_all	weather_id	weather_main	weather_description	weather_icon
2	2015-01-01 00:00:00+01:00	Valencia	270.475	270.475	270.475	1001	77	1	62	0	0	0	0	800	clear	sky is clear	01n
3	2015-01-01 01:00:00+01:00	Valencia	270.475	270.475	270.475	1001	77	1	62	0	0	0	0	800	clear	sky is clear	01n
4	2015-01-01 02:00:00+01:00	Valencia	269.686	269.686	269.686	1002	78	0	23	0	0	0	0	800	clear	sky is clear	01n
5	2015-01-01 03:00:00+01:00	Valencia	269.686	269.686	269.686	1002	78	0	23	0	0	0	0	800	clear	sky is clear	01n
6	2015-01-01 04:00:00+01:00	Valencia	269.686	269.686	269.686	1002	78	0	23	0	0	0	0	800	clear	sky is clear	01n

The energy dataset contains 29 columns of data:

- Datetime index localized to CET
- Energy Demand in unit MW: biomass generation, coal/lignite generation, coal gas generation, gas generation, coal generation, oil generation, shale oil generation, peat generation, geothermal generation, hydro1 generation, hydro2 generation, hydro3 generation, hydro4 generation, sea generation, nuclear generation,other generation, other renewable generation, solar generation, waste generation, wind offshore generation, wind onshore generation, forecasted solar generation, forecasted offshore wind generation, forecasted onshore wind generation, forecasted electrical demand, actual electrical demand
- Cost per energy in unit EUR/MWh: forecasted price, actual price.

Table 2. Sample data rows from the energy dataset.

1	time	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	generatio	forecast s	forecast w	forecast w	total load	total load	price day	price actual
2	2015-01-0	447	329	0	4844	4821	162	0	0	0		863	1051	1899	0	7096	43	73	49	196	0	6378	17		6436	26118	25385	50.1	65.41
3	2015-01-0	449	328	0	5196	4755	158	0	0	0		920	1009	1658	0	7096	43	71	50	195	0	5890	16		5856	24934	24382	48.1	64.92
4	2015-01-0	448	323	0	4857	4581	157	0	0	0		1164	973	1371	0	7099	43	73	50	196	0	5461	8		5454	23515	22734	47.33	64.48
5	2015-01-0	438	254	0	4314	4131	160	0	0	0		1503	949	779	0	7098	43	75	50	191	0	5238	2		5151	22642	21286	42.27	59.32
6	2015-01-0	428	187	0	4130	3840	156	0	0	0		1826	953	720	0	7097	43	74	42	189	0	4935	9		4861	21785	20264	38.41	56.04

Independent variables: weather measurements, cities, energy generation/types, time
Dependent variables: electrical demand/consumption, prices/cost

We hypothesize weather measurements and time will be the most important variables for predicting electrical demand and prices/cost.

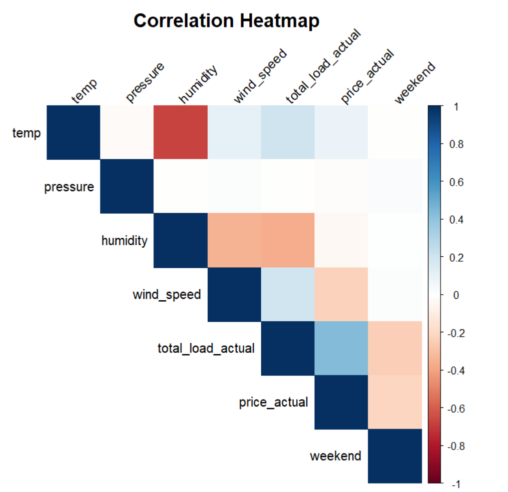

Figure 1: Correlation heatmap of data features.

DATA CLEANING

Preparing the datasets for analysis involved a series of steps in R. The following steps summarize the process taken to clean the energy and weather datasets:

1. **Uniform Column Names**: All column names were made consistent across the dataset for increased clarity and ease of data manipulation.
2. **Adding Time Details**: Columns for dates and time in a date format (such as year, month, and season) were added to more easily see trends over time.
3. **Data Pruning**: Redundant and non-informative columns were removed since they did not provide useful information to the analysis.

4. **Weather Data Consolidation**: Meteorological data points were averaged, providing a macroscopic view of national weather patterns.
5. **Dataset Integration**: The weather and energy datasets were joined by time columns, creating a singular dataset used for analysis (35,064 rows).
6. **Excluding Incomplete Data**: Incomplete records (47 out of 35,064 rows) were removed to prevent skewing the analysis results.
7. **Quality Control Checks**: A final review was conducted to review the dataset, looking for gaps or missing information.

Figure 1 displayed the correlation heatmap of the cleaned dataset. There is a negative correlation between : load and humidity, price and wind speed, load and weekend, price and weekend. There is a positive correlation between: load and wind speed, load and price.

DATA MODELING

Predicting Load - Random Forest

We are primarily interested in predicting the electric load over the timeframe of one day. This falls in the domain of *Short Term Load Forecasting* (STLF). Kusher et al. describe that the majority of studies for load prediction on this timeframe have implemented time series methods like ARIMA or artificial neural networks (ANNs). A major flaw of time series analyses is that they tend to exclude external variables which may affect performance.[1] Azeem et al., report that there are ten main components that models should consider: Time, Meteorology, Application, Geography, Economic factors, Historical information, Events, Data quality, Technology, and Distributed power resources [2]. Of those factors, meteorology is denoted as being one of the most dominant so its exclusion will naturally lead to high errors when considering a large-scale geographic domain like Spain. ANNs on the other hand, have the benefit of learning the intricate relationships between an arbitrary number of features. However, it's challenging to discern feature importances from them which is a fundamental component of this study in order to prescribe solutions to the utility company. We therefore propose a random forest as a middle ground method, as they are able to model complex relationships while offering a native means to interpret feature significance.

Of the ten components reported by Azeem, our dataset allows us to directly consider time (hour, weekend), meteorology, and economic factors (electricity price) as direct predictors. Seasonal patterns have been noted to be important in the literature [1][2] so we additionally incorporate that directly as a factor. We also directly consider historical information and data quality in our training data. The meteorology category is really a group of numerous sensible weather observations including irradiance, rainfall, snowfall, temperature, wind speed, and humidity. There are mixed results regarding the importance of some of these factors depending on the region and application [1] so we include them all anyway. We incorporate both hourly and three hourly rain and snow measurements as predictors even though they have a degree of covariance. We expect consumer behavior to vary between a passing thunderstorm versus an extended period of rain, for example.

Predicting Price - Linear Regression

To model the price response variable, we first started with a baseline model to compare. To do this we split our modeling into three different "baseline" models, a multiple linear regression with all of the training data (80%), a multiple linear regression with potential outliers removed, and a multiple linear regression with outliers removed and the response transformed via the optimal lambda in the "box cox" method. We will refer to these 3 as full model 1, full model 2, and full model 3, respectively. After fitting full model 1, we decided to use the cook's distance to identify potential outliers in the training data and remove them. After removing potential outliers, we then fit full model 2. Finally, to evaluate multicollinearity we checked the VIF to see if there were any features that had a VIF score greater than or equal to 10. Noticing some features that met this criteria, we decided that maybe a transformation was in order. With this thought, we used the *boxcox* method from the *mass* package to obtain the optimal lambda for transformation and use it to fit full model 3, transforming the price_actual response variable. After fitting the 3 baseline models, we decided that we needed to implement some residual analysis to make sure that regression assumptions hold. Below are some residual visualizations for full model 1 to help us check this. Please note, we did this for model 2 and model 3 as well.

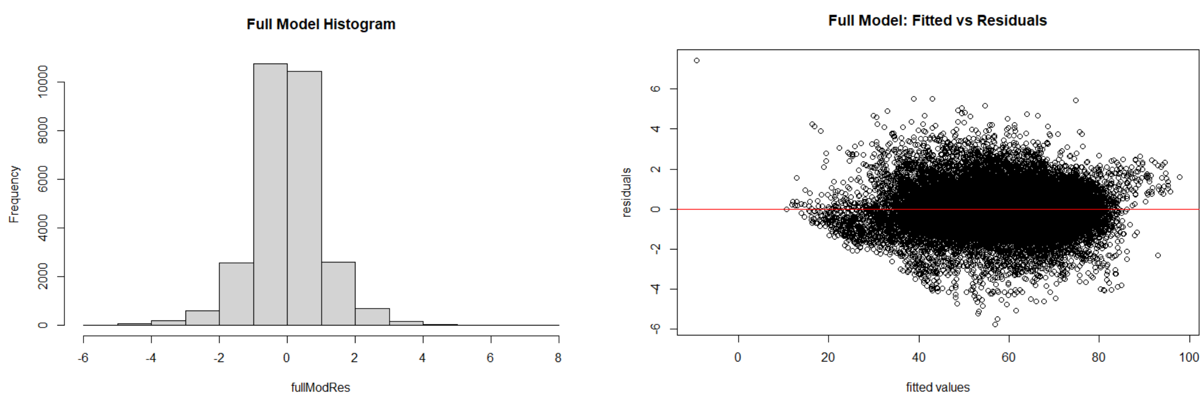

Figure 2*: Residual Analysis of Full Model 1*

Finally, after checking the assumptions, we used the testing dataset (20%) to find the average RMSE of each of these full "baseline" models via monte carlo cross validation with 100 iterations.

To assist with variable selection, we tested stepwise, LASSO, and elastic net approaches. For the stepwise model, we used a forward stepwise regression. We used Bayesian Information Criterion (BIC) for this model to help balance fitness and complexity of the model. [3] With this approach, the predictors of day, generation_fossil_gas, generation_fossil_oil, generation_solar, pressure, price_day_ahead, rain_1h, rain_3h, temp_max, total_load_actual, and year were removed. The LASSO model removed the temp and pressure predictors, and the elastic net model removed only the temp predictor. We then created 3 new models using the variables selected from the stepwise, LASSO, and elastic net regressions respectively, to be compared against each other. We performed residual analysis on all models to evaluate key indicators of ‘ linearity, normality, and heteroscedasticity and assess the validity of each model, with each of the models passing the same assumptions discussed above except for the stepwise model. The qqplot testing residuals vs. fitted values in this model showed some signs of variance.

Next we evaluated the best of the 4 models using ANOVA and compared the selected model’s predictions to the testing data by calculating the average Root Mean Square Error (RMSE) of each. The end result will be an evaluation of all predictors that have a statistical relationship with price to be used in the utility company’s energy price forecasting.

MODELING RESULTS

Price Prediction

In regards to the results of our analysis for predicting price through linear regression, we found through looking at RMSE and ANOVA analysis that the elastic net regression model performed the best on the testing data (after cross validation). Please see the table below with the RMSE values for all of the models tested:

Table 3: RMSE Results

Model:	*Full*	*Stepwise*	*LASSO*	*Elastic Net*
RMSE:	7.756809	7.816911	7.756053	7.756156

As we can see from the table above, all of the models were pretty close to each other in terms of the RMSE testing error, but the best (by a small margin) was LASSO, beating the Elastic Net Model by only 0.000103. The Stepwise model performed the worst with an RMSE of 7.816911. However, in the ANOVA analysis the stepwise and elastic net models demonstrated better P-values (at $< 2.2e-16$ and 0.02903 respectively). Because the stepwise model showed signs of high variance in our residual analysis, we conclude it may be overfit to the data. Based on these results, we will go with the elastic net model as our “final selection” model to predict the actual price. We prefer the elastic net model because the LASSO and elastic net models had very similar RMSE values, but the elastic net model performed much better on the ANOVA analysis.

Load Prediction

The dataset was partitioned into training and test by assigning all of the 2015 data to the training dataset and all of the 2016 data to the test dataset. Rows with empty values were omitted. A baseline model was trained using all the named factors above (hour, weekend, season, price_actual, temp, humidity, wind_speed, rain_1h, rain_3h, snow_3h, clouds_all) with default parameters in R's *randomForest* package. These defaults include 500 decision trees and *sqrt*(11) features to consider at each split (*mtry*). Mtry was initialized as the square root of the number of columns in the dataframe. We used Out-of-Bag error to gauge convergence in the number of trees, which as denoted in Figure 1, occurs at around 200 trees. The mean squared of the residuals for this baseline model was 4765023 and percent variance explained was 77.89. The out of bag error plot shown in Figure 2 shows the optimal number of trees is approximately 200. Out of bag error measures how well the random forest model predicts on the data points that were not used to train each individual tree. It provides an estimate of the model's performance on unseen data without the need for a separate validation set.

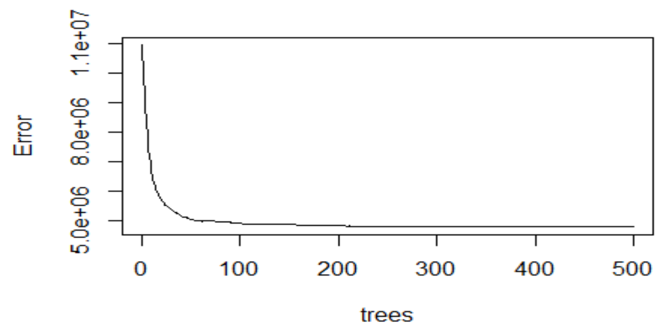

***Figure 3**: Out-of-bag error plotted against number of trees.*

The *tuneRF* function was implemented to tune the model's *mtry*, which refers to how many features are randomly chosen for consideration at each split. Mtry can be set to a fixed number or a proportion of the total number of features. A smaller mtry value makes the trees more diverse and less correlated, but it might lead to less predictive power. A larger mtry value increases the correlation between trees but might result in more accurate predictions.It tries successive values of mtry starting with the default value, then from the closest integer to the default value, iterates by 2, and stops when relative OOB error improvement is less than 1% from the best value tried. As the plot in Figure 3 shows, the optimal value is achieved at *mtry* = 3 with an out of bag error of 4803164.

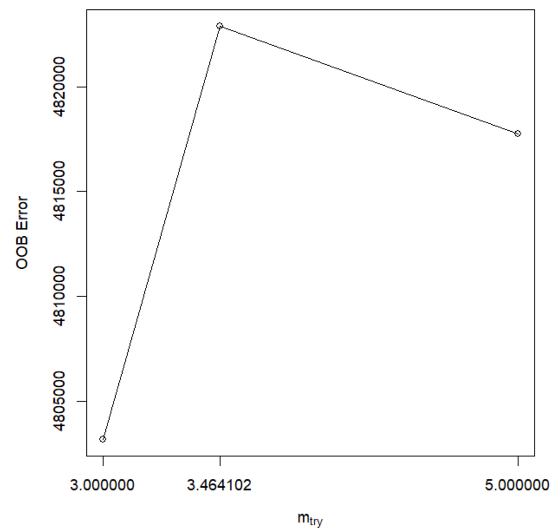

***Figure 4**: Out-of-bag error plotted against mtry, the number of features to consider at each decision tree split.*

The variance importance plot shown in Figure 4 displays hour, price_actual, and humidity are the three most important features in the base model. Temperature is only marginally less important than humidity.

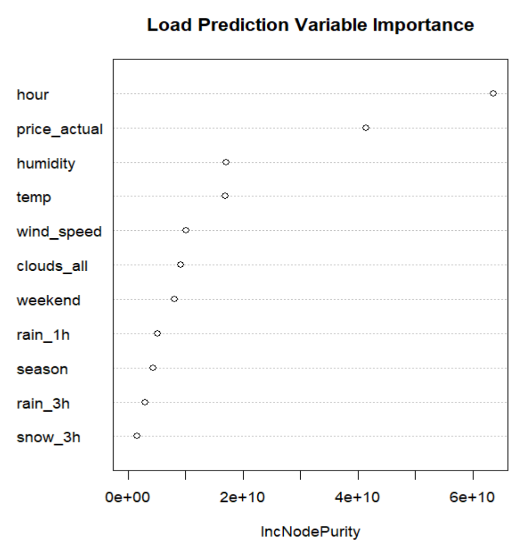

***Figure 5**: Feature importances based on change in error upon feature removal*

Snow_3h, rain_3h, and season are the least important. A variance important plot shows the importance of each variable (feature) in making predictions with the Random Forest.

The caret package in R was used to tune *mtry* more thoroughly with 5-fold cross validation using 200 trees. Mtry of 7 had the lowest root mean square error, highest R squared or percent variability explained, and lowest mean absolute error. The most important features for this cross-validated were determined to be hour, price_actual, and temp. Humidity is only marginally of lower importance than temperature. The majority of these predictors align with the baseline random forest model. Figure 5 displays the corresponding RMSE plot for the *mtry* tuning.

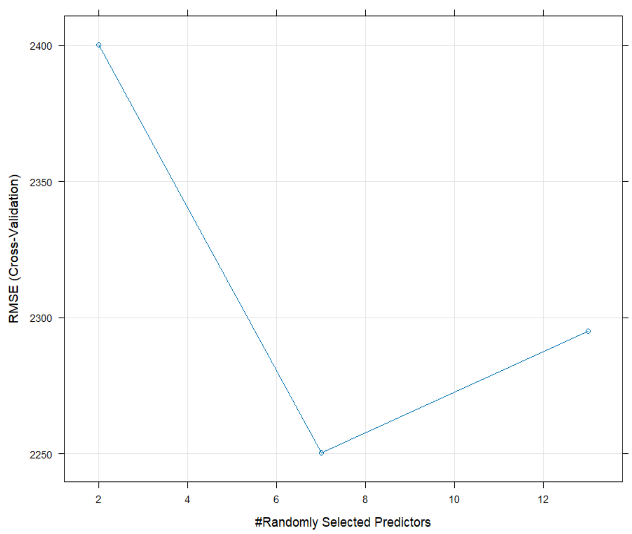

Figure 6*: 5-fold CV RMSE vs mtry, the number of randomly selected features at each decision tree split*

Finally, we tuned the model for optimal features via 5-fold cross validation using carat's recursive feature selection algorithm. Cross validation found that the best accuracy was achieved with all features. It makes sense that the wind and precipitation factors contributed comparatively little to model variance while also being important enough to yield improved accuracy when considered by the random forest. Spain has a warm mediterranean climate with dry summers. Windy days and precipitation are scarce in the warmer months so for much of the year they are not relevant to electricity demand. When rain does fall in the summer, it's usually brief and in the form of localized thunderstorms. But when winter storms affect the country, precipitation is often widespread and of longer duration, often accompanied by breezy winds. It would make sense that residential electricity demand is greater on such days as people stay home. Although such weather events constitute a very small proportion of our dataset, Spanish utility managers ought not discount them.

We made predictions using the baseline random forest model with 500 trees and sqrt(11) *mtry*, the tuned model with 200 trees and sqrt(11) *mtry,* the model with 200 trees and 3 *mtry,* and the tuned model with 200 trees and 7 *mtry* to compare predictions for total actual load for the unseen Jan-Feb 2016 test data. Error results are presented below in Table 4 for these two models and for intermediate tuning steps.

The tuned model yielded only small improvements to the error compared to the baseline and there was no improvement in building a deeper tree by considering 7 features at each split compared to 3 The scale of the load is on the order of ~40000 MW so Mean Absolute Percent Error (MAPE) is a more useful metric than either Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). Both models tuned for *mtry* yield predictions that normally deviate by 8.24% from the observed load.

Table 4: Error metrics for various iterations of the random forest model

Model	MSE	RMSE	MAPE
Baseline model (sqrt(11) mtry and 500 trees)	9817797	3133	0.0826
Tuned model (sqrt(11) mtry and 200 trees)	10118479	3181	0.0838
Cross validated RF (7 mtry and 200 trees)	9756860	3124	0.0824
Cross validated RF (3 mtry and 200 trees)	9750226	3123	0.0824

CONCLUSION

Using the final Elastic Net regression model, we were able to generate a predicted price that deviated on average by $7.76 from the actual price. In regards to pricing, temp and pressure do not play a significant role in determining price when using the model. All of the models performed similarly in regards to the RMSE metric, but after the ANOVA analysis we saw that the elastic net outperformed the rest (leaving out stepwise due to overfitting concerns). This is within the range of acceptable accuracy that would be useful for residential and industrial price prediction of factor influences on their energy prices. This analysis could be useful by different entities in terms of budgeting energy consumption as well as optimization of production.

We were able to successfully demonstrate the viability of electric grid load prediction for winter months using a random forest model. Important features for prediction are hour, price_actual, temp, humidity and a categorical variable denoting whether the forecasted day is a weekend. Load managers need to consider these features at a minimum but precipitation features offer improved predictions when they are incorporated, especially in the winter when storms are more frequent. Using 2015 data as the training dataset and Jan - Feb 2016 data as the testing dataset, the random forest performance peaks at 200 trees and three features considered at each split. There is no advantage in implementing deeper trees. Model predictions deviate by about 8%. Utility companies can improve the model further by incorporating additional features related to the grid and we recommend exploring load for different sectors like residential or industrial for more targeted predictions. At a minimum though, we believe our model can be used to guide decisions regarding anticipated electric production, fuel purchases, overhead, and strategies to mitigate potential stress on the electric grid.

REFERENCES

[1] Kuster, C., Rezgui, Y., & Mourshed, M. (2017). Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society*, *35*, 257–270. https://doi.org/10.1016/j.scs.2017.08.009

[2] *Electrical Load Forecasting Models for Different Generation Modalities: A review*. (2021). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9576703

[3] Lu, H., Ma, X., Ma, M., & Zhu, S. (2021). Energy price prediction using data-driven models: A decade review. *Computer Science Review*, *39*, 100356. https://doi.org/10.1016/j.cosrev.2020.100356